

Old English UD with Limited Supervision: Character-Aware Stanza and spaCy with In- Domain tok2vec Initialisation

Ana Elvira Ojanguren López

Universidad de La Rioja, Logroño, Spain

Dario Metola Rodríguez

Universidad de La Rioja, Logroño, Spain

Javier Martín Arista

Universidad de La Rioja, Logroño, Spain

Abstract

Old English presents distinctive NLP challenges: supervision is scarce, spellings vary across sources, and inflection is rich. Within the Universal Dependencies framework (de Marneffe et al., 2021), we contrast two practical routes: character-aware Stanza stacks paired with a bilinear arc-scoring parser (Dozat & Manning, 2017; Straka & Straková, 2019) and spaCy pipelines trained either from scratch or with in-domain tok2vec initialisation (Bojanowski et al., 2017).

We report tokenisation, sentence segmentation, POS and morphological features, lemmatisation, and dependency parsing on matched splits. Stanza is evaluated at c.25k and 50k supervised tokens; spaCy at ~20k under three settings: scratch, tok2vec-initialised, and a compact transformer without extensive self-supervised pretraining (He et al., 2020). Results are contextualised against multilingual transfer (Villa & Giarda, 2023) and recent Old English UD modelling (Martín Arista et al., 2025a; 2025b).

Doubling supervision for Stanza yields an LAS increase from 77.33 to 86.15, with corresponding gains in UAS, indicating that character-centric representations profit from additional examples (Kondratyuk & Straka, 2019). At 20k, spaCy benefits markedly from in- domain subword pretraining: relative to the scratch model, UAS improves from 78.26 to

83.24 and LAS from 68.10 to 74.23, alongside stronger POS and morphology. By contrast, a scratch compact transformer underperforms (LAS 45.51; sentence F1 40.07), underscoring the need for substantial self-supervision (He et al., 2020).

Fine-grained tags and lemma normalisation remain brittle and non-monotonic across regimes, echoing known issues in historical text normalisation and inventory alignment (van der Goot et al., 2020). Overall, improvements separate into gains from added supervision (Stanza) and gains from better initial representations (spaCy), while cross-lingual transfer remains below in-domain, character-aware models for structural accuracy (Villa & Giarda, 2023).

Practically, choose Stanza when dependency quality is paramount and more annotation is available; prefer spaCy with tok2vec initialisation when efficiency and robust tagging are priorities.